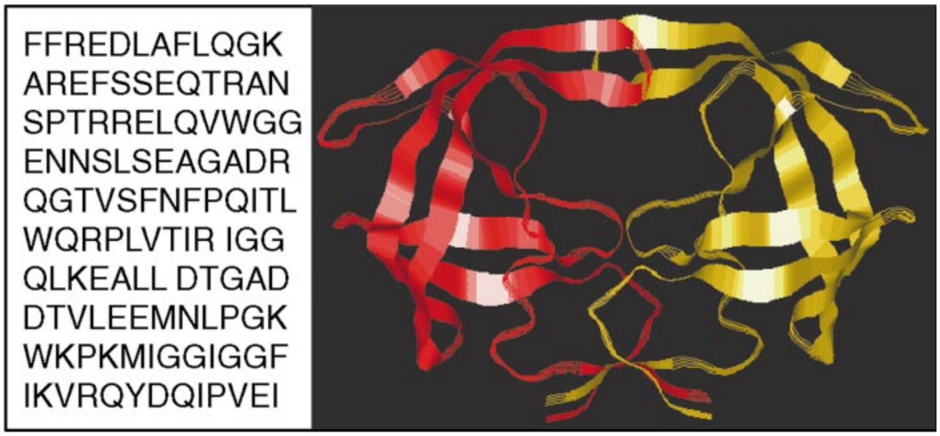
University of California
Lawrence Livermore
National Laboratory

DataFoundry

Data Warehousing and Integration for Scientific Data Management



A partial sequence of the Human Immunodeficiency Virus protein (HIV Protease) is shown on the left, and the protein's 3-D structure is shown on the right.

Technology:

DataFoundry is an advanced data warehouse that combines leading-edge technology from the machine learning and database communities to enable powerful ad-hoc query and inference access to integrated data from multiple independent data sources representing different but related conceptual domains.

Applications:

DataFoundry is relevant to any application area requiring the use and sharing of large-scale, distributed, heterogeneous information. Currently, DataFoundry is being developed in the genomics arena to enable new computational analysis techniques for functional genomics and structural biology.

Computational methodologies are proving to be a viable and cost-effective alternative to the slower and more expensive experimental methods prevalent in many scientific domains. These new techniques are generating large amounts of highly complex and dynamic data which must be adroitly managed for effective analysis. Scientific data management is a critical enabling technology that addresses this need.

Functional genomics is a domain of great practical importance that is currently facing an explosive growth in the complexity and availability of certain types of data. Functional genomics involves identifying the role encoded proteins play in an organism. This work will have an immense impact on the overall health and well-being of humanity over the next century. This is not just due to the medical advances in treating human illness that will be made possible, but also because of the introduction of more disease resistant and prolific varieties of wheat, rice, cattle and other economically vital species.

Determining the structure of each of the naturally occurring proteins would greatly facilitate the goal of functionally characterizing genomic sequences. To date, the primary techniques for identifying a protein's 3D structure rely on NMR or crystallography. However, without massive additional outlays for new equipment and labs, they can determine only 100–200 unique protein structures per year. In comparison, over the next five years, the human genome project alone will sequence on the order of 100,000 proteins. Bridging this enormous gap in capability represents a huge challenge to the scientific community.

To meet this challenge, current experimental techniques need to be supplemented with a strong computational approach which can identify the 3D structure of a protein based on an analysis of the 1D sequence data. One of the essential cornerstones of an effective computational analysis program is that:

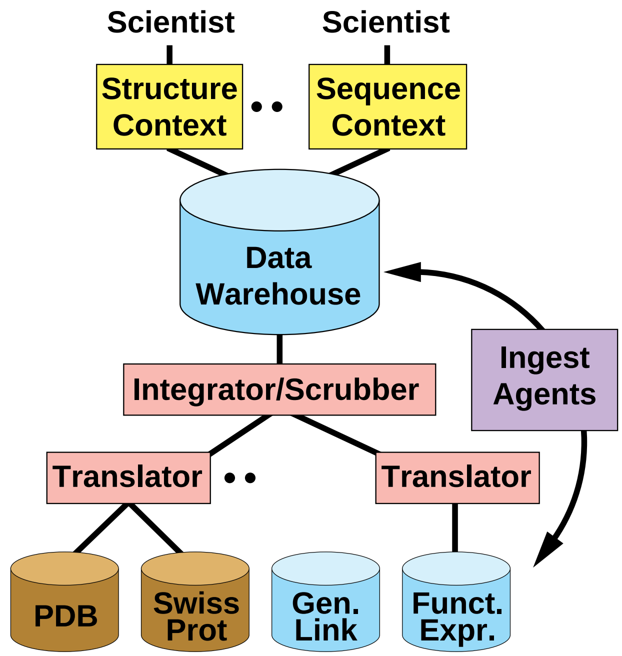
Scientists must be able to access and analyze data from sources in multiple conceptual domains through an integrated interface.

Difficulties in Managing Scientific Data

Data warehousing and integration technologies play a central enabling role in addressing this challenge because they are geared towards integrating distributed, heterogeneous databases. Unfortunately, the warehouse solutions available today do not scale well to dynamic, scientific environments. In these environments, data from multiple independent data sources representing different domains must be integrated. For example, in functional genomics the scientist needs access to data from multiple areas including DNA sequence, protein structure, genetic trait, genetic linkage, physical mapping, and taxonomy. The primary characteristic of scientific data is that the data and the conceptualization of the data is highly dynamic. Managing data which is continually evolving raises several significant data management issues which current warehouse solutions are too rigid and labor-intensive to adapt to. Because the scientist's conceptualization of the data is highly dynamic, the underlying schema of a data source tends to be modified frequently. This alone makes the cost and effort required to maintain a functional, conventional warehouse prohibitive. Additional complexities are that the schema of each data source is radically different, and may represent data in a variety of formats from flat files to OODBMSs. The nomenclature in a single data source (not to mention different data sources) is often non-standard and conflicting. The data can be duplicative, erroneous, and conflicting. Finally, data must be frequently updated, or the warehouse will grow obsolete.

The DataFoundry Approach

The DataFoundry approach is based on the mediated warehouse architecture shown to the right, which is designed to provide a structured, modular framework for integrating multiple independent data sources. Our approach to managing living data is multi-pronged. First, a metadata database that directs automated warehouse rebuilds is used, thus insulating the internal warehouse schema and code from changes occurring in the external data sites. Second, data mining techniques are used to automate the object-identification problem faced in any integration project. Finally, the DataFoundry internal schema is a simple but powerful model. It is comprised of three components: a set of shared objects, relationships between these objects, and an ontology for the data source schemata. A shared object is a unifying concept composed of information from multiple data sources. The objects are associated with each other through both direct and derived relationships. The derived relationships are based on a new service that permits unique inference and analysis based on computing and providing access to multiple homology calculations. The third part of the model, the ontology, is used by the ingest agents to reduce the user interaction required to handle schema changes.



The DataFoundry Mediated Data Warehouse Architecture

Enhancing user efficiency and capability are the primary goals when building effective warehouses for large data; this has both computational and ease-of-use implications. Current warehouse solutions include monolithic centralized storehouses, or approaches based in multidatabase or federated database techniques. These approaches differ in how two fundamental warehousing questions are answered: (1) how much data is stored locally in the warehouse, and (2) how detailed is the warehouse schema. For a variety of reasons ranging from inflexibility to change, to heavy user burdens for accessing the underlying data, to performance that is heavily dependent on network and data source availability, none of these approaches are adequate for scientific data. DataFoundry is a combination of the above approaches. The schema is detailed enough to relieve the user burden, and contains information to help automate schema integration. Data is kept both locally and in the external sources. Decisions on which data to keep local is made based on maximum expected utility - or minimizing the query response times to common and important queries. Queries involving external data will be automatically translated into the external data site format before being dispatched. Data ingest will be based on intelligent agent technology with change detection techniques, resulting in a continuous and incremental update.

DataFoundry is designed to be flexible enough to scale to the demands of realistic scientific environments, while providing a powerful query and inference capability across a wealth of data. For example, the following query combines information from five genetics data sources (distinguished by color) into a single, intuitive request:

> **Compare the structural homologies of all proteins expressed in the mouse eyeball after 45 days which lie on on chromosome 7, 8, or 11 and are implicated with glaucoma.**

Multidisciplinary Collaboration

The DataFoundry project is a multidisciplinary effort involving scientists from the Center for Applied Scientific computing and the Biology and Biotechnology Research Program.

For additional information about the DataFoundry project, contact Ron Musick, 510-424-5015, rmusick@llnl.gov.

Work performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48